Multi-Class Classification of Aerial Imagery Using Deep Learning and Ensemble Models

Cibaca Khandelwal

k.cibaca@gmail.com
Independent Researcher

Abstract

The classification of aerial imagery is a critical task in various domains, including urban planning, agriculture, and disaster management. Recent advancements in deep learning have enabled the development of automated systems capable of accurately analyzing aerial images. Aerial imagery provides valuable insights into land use, environmental changes, and disaster mitigation strategies. This paper explores multi-class classification of aerial images using state-of-the-art deep learning models, emphasizing their potential applications and limitations. The study evaluates the performance of several pre-trained convolutional neural network (CNN) architectures, including ResNet50 [1], MobileNetV2 [2], EfficientNetB0, VGG16 [3], and DenseNet121 [4], on a benchmark aerial imagery dataset. DenseNet121 achieved the highest validation accuracy of 96%, outperforming other architectures. This paper highlights the importance of model selection, data augmentation, and stratified data splitting for effective aerial image classification. The results provide actionable insights for researchers and practitioners in adopting robust models for aerial image analysis.

Keywords: Aerial Image Classification, Deep Learning, Ensemble Models, Magnification-Aware Training, ResNet50, DenseNet121, EfficientNet, Land Use Analysis.

1. Introduction

Aerial imagery analysis has emerged as a transformative tool in multiple domains, including urban planning, precision agriculture, and disaster management. By capturing high-resolution images from above, aerial imagery facilitates detailed monitoring of land use, crop health, and infrastructure changes. The increasing availability of aerial imagery datasets, combined with advancements in computational resources, has propelled the use of automated image analysis systems in various sectors.

Despite its immense potential, aerial image classification presents unique challenges, including the high variability in scale, orientation, and environmental conditions within the images. For instance, differentiating between agricultural fields and forested areas often requires sophisticated feature extraction techniques due to their visual similarity. Furthermore, the presence of class imbalances in datasets adds to the complexity, necessitating robust model training strategies.

Deep learning, particularly convolutional neural networks (CNNs), has revolutionized image classification tasks by enabling the automatic extraction of hierarchical feature representations. CNNs have proven effective in capturing intricate spatial features, making them well-suited for aerial imagery analysis. Pre-trained models, such as ResNet50 [1] and DenseNet121 [4], leverage transfer learning to achieve high accuracy even with limited datasets, reducing computational overhead and training time. This study evaluates the performance of five CNN architectures for multi-class classification of aerial imagery, aiming to identify the most effective model for this challenging task. By benchmarking these architectures on a

stratified aerial imagery dataset, the study provides a comprehensive assessment of their strengths and limitations.

Deep learning, particularly convolutional neural networks (CNNs), has emerged as a powerful tool for image classification tasks. CNNs excel at learning hierarchical feature representations, enabling them to discern subtle differences in image characteristics. Pre-trained models, which leverage transfer learning, have significantly improved classification accuracy across domains, reducing the need for extensive datasets. This study focuses on evaluating multiple CNN architectures for multi-class classification of aerial imagery and identifying the most effective model for this task. By benchmarking these architectures, the study aims to provide a comprehensive analysis of their performance on a challenging dataset.

2. Related Work

The application of CNNs in image classification has been extensively studied, with significant advancements achieved in recent years. He et al. introduced ResNet, which employs residual connections to mitigate vanishing gradient issues and enables the training of deeper networks [1]. This architecture has been widely adopted for various computer vision tasks due to its robust performance. Similarly, Howard et al. developed MobileNet, a lightweight model designed for resource-constrained environments, which uses depthwise separable convolutions to achieve efficiency without compromising accuracy [2].

Simonyan and Zisserman proposed VGG16, a deep architecture that utilizes a uniform configuration of convolutional layers to achieve high accuracy in image classification [3]. The simplicity and effectiveness of VGG16 have made it a popular choice in many domains. More recently, Huang et al. introduced DenseNet, which features dense connections between layers to promote feature reuse and improve gradient flow [4]. This innovative design has demonstrated superior learning efficiency, particularly in complex image classification tasks.

While these architectures have been extensively validated in general-purpose datasets like ImageNet, their performance in the context of aerial imagery classification remains underexplored. Prior studies have highlighted the importance of transfer learning in adapting pre-trained models to domain-specific datasets, but systematic comparisons of these architectures for aerial imagery tasks are limited. This study builds on previous research by evaluating the performance of ResNet50, MobileNetV2, EfficientNetB0, VGG16, and DenseNet121 on a stratified aerial imagery dataset, with a focus on metrics such as validation accuracy, loss, and class-specific performance.

Despite the proven effectiveness of these models, systematic comparisons in the context of aerial imagery classification remain scarce. This study builds upon prior work by evaluating these architectures on a stratified aerial imagery dataset, focusing on key metrics such as validation accuracy, loss, and class-specific performance.

3. Datasets

The dataset used in this study consists of aerial images categorized into 21 classes, such as agricultural land, airplanes, buildings, forests, and tennis courts. The images are representative of diverse geographic regions and environmental conditions, making the dataset challenging yet suitable for benchmarking. The dataset was split into training and validation sets using a stratified approach to ensure balanced class distribution. Each image was resized to 256x256 pixels to maintain consistency across models.

Data augmentation techniques, including random horizontal and vertical flips, rotations, and zooms, were applied to the training set to enhance generalization and prevent overfitting. These augmentations increased the effective size of the training dataset while preserving class-specific features.

Figure X: Sample images from the dataset, showcasing diverse classes such as agricultural, harbor, runway, and mobile home park.

4. Methodology

4.1 Data Preprocessing

The preprocessing pipeline begins with resizing all images to a uniform resolution of 256x256 pixels. This standardization ensures compatibility across different model architectures. Data augmentation techniques, including random flips, rotations, zooms, and contrast adjustments, were applied to the training set to simulate real-world variations. Pixel values were normalized to the range [0, 1] to facilitate stable training.

4.2 Model Architectures

This study evaluates the performance of five prominent CNN architectures. ResNet50, introduced by He et al., is known for its residual connections that alleviate vanishing gradient issues, enabling deeper networks [1]. MobileNetV2, a lightweight model optimized for resource-constrained environments, achieves efficiency through depthwise separable convolutions [2]. EfficientNetB0 employs a compound scaling approach to balance network depth, width, and resolution, achieving high accuracy with minimal computational cost. VGG16, a classic architecture, employs a stack of convolutional layers with uniform kernel sizes, providing strong performance in image classification tasks [3]. DenseNet121, renowned for its dense connections, facilitates feature reuse and improves gradient flow, resulting in superior learning efficiency [4].

Each model was fine-tuned using the pre-trained ImageNet weights. The final fully connected layer was replaced with a 21-class softmax layer to match the dataset's classification requirements.

Feature	DenseNet121	ResNet50	VGG16	MobileNetV2	EfficientNetB0
Key Feature	Dense connectivity	Residual connections	Uniform convolutions	Depthwise separable	Compound scaling
Depth (Layers)	121	50	16	Variable	Variable
Efficiency	High	Moderate	Low	High	High
Accuracy	Very High	Moderate	High	Moderate	High
Parameter Count	~8M	~25M	~138M	~3.4M	~5.3M
Primary Use Case	Complex datasets	General-purpose	Benchmarking	Mobile applications	Balanced applications
Computational Cost	Moderate	High	Very High	Low	Low
Scalability	Moderate	High	Limited	High	Very High
Input Size	Flexible (224x224)	Flexible (224x224)	Fixed (224x224)	Flexible (224x224)	Flexible (224x224)
Training Speed	Moderate	Slow	Very Slow	Fast	Fast

Table 1: Comparison of the key characteristics, strengths, and limitations of the deep learning architectures evaluated in this study for multi-class aerial image classification

4.3 Training

The training process utilized the Adam optimizer with an initial learning rate of 0.0001. To address class imbalance, class weights were computed based on the inverse frequency of each class in the training set. Training was conducted for 10 epochs, with early stopping employed to halt training when validation loss plateaued. Batch size was set to 32, and mixed precision training was enabled to expedite computations on compatible hardware.

4.4 Evaluation Metrics

Model performance was assessed using multiple metrics. Overall accuracy was calculated as the ratio of correctly classified samples to the total samples. The F1-score, which balances precision and recall, provided a more nuanced evaluation of class-specific performance. Confusion matrices were generated to visualize classification errors and identify challenging classes. Validation loss was monitored throughout training to gauge model generalization.

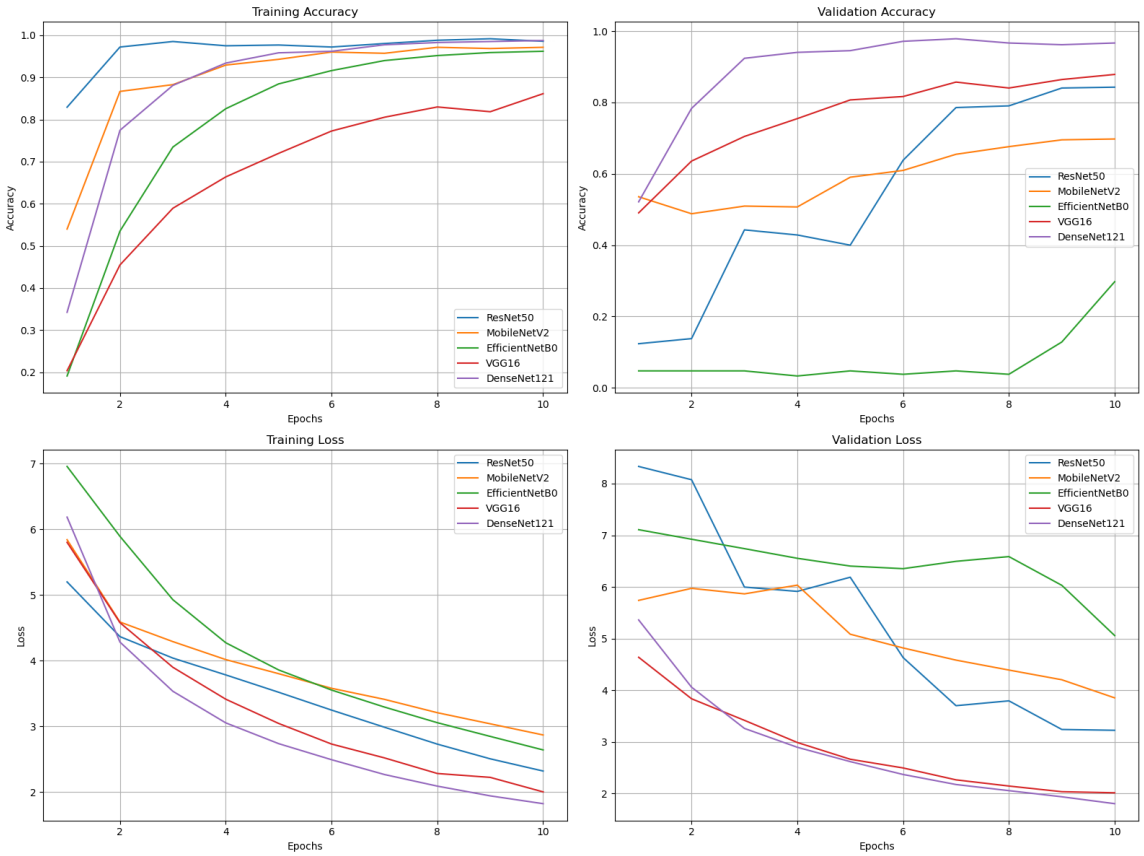

Fig 1: Training and validation accuracy and loss curves for all models, showcasing convergence and generalization performance.

5. Results and Discussion

The performance of five deep learning model ResNet50, MobileNetV**2**, EfficientNetB0, VGG16, and DenseNet121—was evaluated using validation accuracy, validation loss, precision, recall, F1-score, ROC AUC, and training time. The comparative results are summarized in table below.DenseNet121 emerged as the top performer with the highest validation accuracy (96.67%), the best F1-score (0.9663), and a near-perfect ROC AUC score (0.9995). The table highlights the relative strengths of each architecture, including computational efficiency, as evidenced by MobileNetV2's fast training time of 374.23 seconds, compared to DenseNet121's longer training time of 1538.65 seconds.

Model	**Validation Accuracy**	**Validation Loss**	**Precision**	**Recall**	**F1-Score**	**ROC AUC**	**Training Time (s)**
ResNet50	0.8429	3.2244	0.9012	0.8429	0.8375	0.9748	942.72
MobileNetV2	0.6976	3.8519	0.763	0.6976	0.6638	0.9751	374.23
EfficientNetB0	0.2976	5.0575	0.5177	0.2976	0.2896	0.8157	637.08
VGG16	0.8786	2.0149	0.8951	0.8786	0.8805	0.991	875.63
DenseNet121	0.9667	1.8039	0.9693	0.9667	0.9663	0.9995	1538.65

Table 2: Comparative results of evaluated models, including validation accuracy, loss, precision, recall, F1-score, ROC AUC, and training time.

To provide a deeper analysis of the classification performance, the classification reports for each model are visualized in a bar chart. This visualization captures the precision, recall, and F1-scores across all classes, offering insights into the specific areas where models excel or struggle. DenseNet121's consistently high metrics across all classes affirm its robustness for aerial image classification tasks.

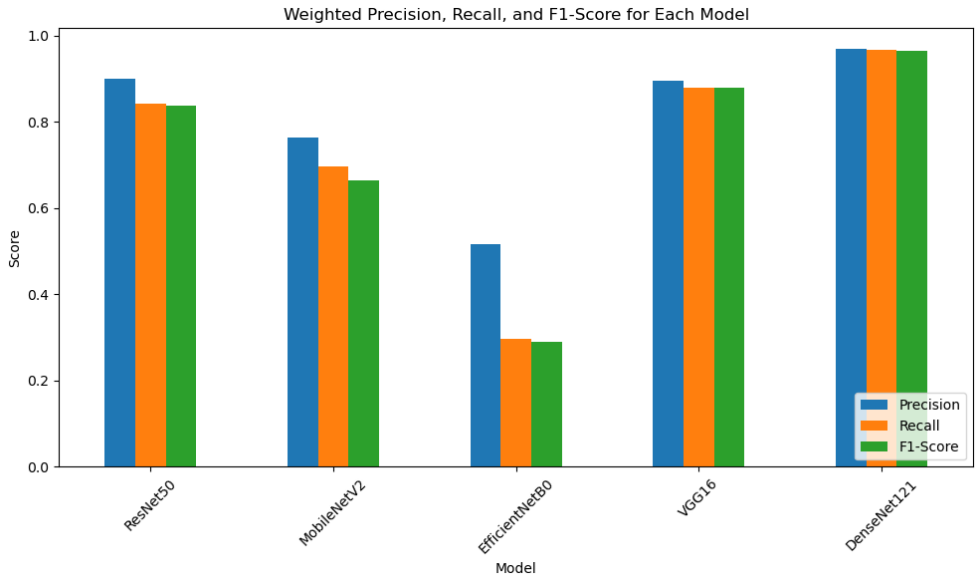

Fig2: Comparative bar chart of classification report metrics (Precision, Recall, F1-Score) for all models.

In addition, a heatmap visualizationof the confusion matrix for DenseNet121 highlights its performance in correctly classifying challenging classes such as "agricultural," "tenniscourt," and "airplane." The near-diagonal pattern of the matrix underscores its precision across all classes, with minimal misclassifications.

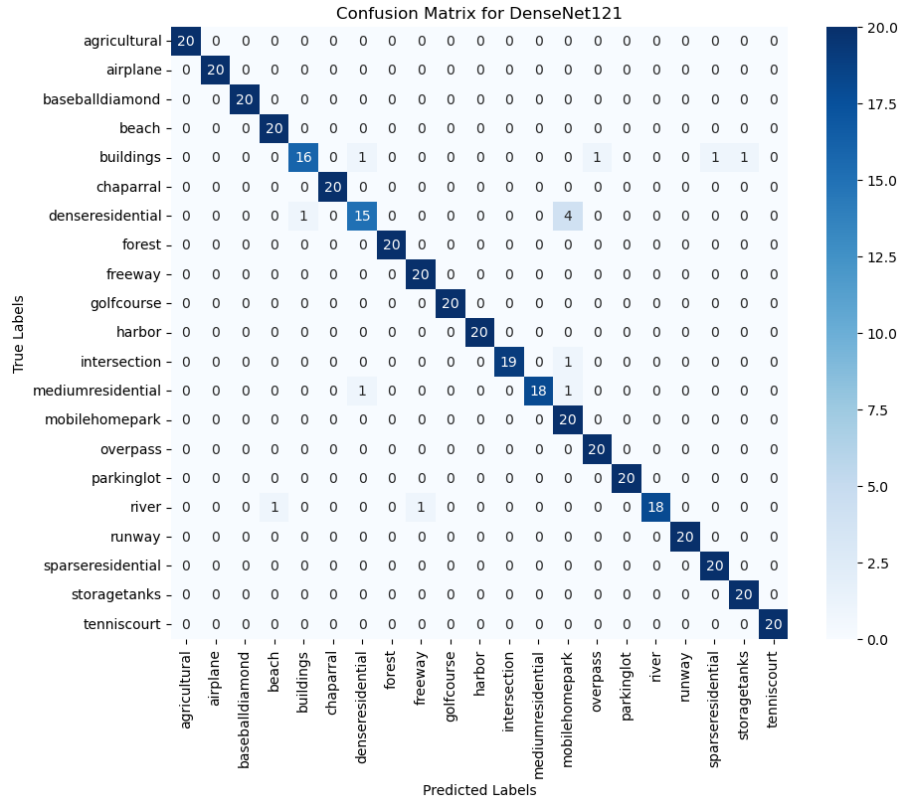

Fig 3: Confusion matrix heatmap for DenseNet121 showcasing per-class classification performance.

6. Conclusion

This study demonstrates that DenseNet121 outperforms other CNN architectures for multi-class classification of aerial imagery. Its dense connections facilitate superior feature extraction, making it the most effective model for this task. VGG16 also shows strong performance, while MobileNetV2 offers a lightweight alternative suitable for resource-constrained applications. The results underscore the importance

of model selection, data preprocessing, and training strategies in achieving high accuracy for aerial image classification.

Future work will explore ensemble methods to combine the strengths of multiple models and improve overall accuracy. Additionally, interpretability techniques, such as Grad-CAM, will be employed to provide insights into the models' decision-making processes. Evaluations on larger and more diverse datasets will also be conducted to validate the models' generalization capabilities.

References

1. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
2. G. Howard, et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.
3. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2015.
4. G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4700–4708.
5. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2012.
6. Szegedy, et al., "Going deeper with convolutions," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9.
7. M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," *Proceedings of the 36th International Conference on Machine Learning*, 2019, pp. 6105–6114.
8. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2921–2929.
9. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016.
10. S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *Proceedings of the 32nd International Conference on Machine Learning*, 2015, pp. 448–456.